

Dictionary Learning Based Sparse Channel Representation and Estimation for FDD Massive MIMO Systems

Yacong Ding, *Student Member, IEEE*, Bhaskar D. Rao, *Fellow, IEEE*

Abstract

Downlink beamforming in FDD Massive MIMO systems is challenging due to the large training and feedback overhead, which is proportional to the number of antennas deployed at the base station, incurred by traditional downlink channel estimation techniques. Leveraging the compressive sensing framework, compressed channel estimation algorithm has been applied to obtain accurate channel estimation with reduced training and feedback overhead, proportional to the sparsity level of the channel. The prerequisite for using compressed channel estimation is the existence of a sparse channel representation. This paper proposes a new sparse channel model based on dictionary learning which adapts to the cell characteristics and promotes a sparse representation. The learned dictionary is able to more robustly and efficiently represent the channel and improve downlink channel estimation accuracy. Furthermore, observing the identical AOA/AOD between the uplink and downlink transmission, a joint uplink and downlink dictionary learning and compressed channel estimation algorithm is proposed to perform downlink channel estimation utilizing information from the simpler uplink training, which further improves downlink channel estimation. Numerical results are presented to show the robustness and efficiency of the proposed dictionary learning based channel model and compressed channel estimation algorithm.

Index Terms

Channel estimation, dictionary learning, compressive sensing, joint dictionary learning, joint sparse recovery, FDD, Massive MIMO

Y. Ding and B. D. Rao are with Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, 92093-0407 USA (e-mail: {yad003, brao}@ucsd.edu).

I. INTRODUCTION

Massive Multiple-Input Multiple-Output (MIMO) systems have been proposed for the next generation of communication systems [1], [2]. By deploying a large antenna array at the base station (BS), the base station is able to perform both receive and transmit beamforming with narrow beams, thus eliminating multiuser interference and thereby increasing the cell throughput. For effective downlink beamforming, it is essential to have accurate knowledge of the downlink channel state information at the transmitter (CSIT). In a time-division duplexing (TDD) system, downlink CSI can be obtained by exploiting uplink/downlink channel reciprocity. The common assumption in Massive MIMO is that each user terminal (UT) only has a small number of antennas and that the BS can use uplink channel information, obtained from the relatively easy uplink training, for downlink beamforming [2]. On the other hand, as frequency-division duplexing (FDD) is generally considered to be more effective for systems with symmetric traffic and delay-sensitive applications, most cellular systems today employ FDD [3], [4]. Channel reciprocity is no longer valid in FDD systems and in order to obtain CSIT, the BS has to perform downlink training. Subsequently, the user needs to estimate, quantize and feedback the channel state information. Downlink training and feedback of CSI in FDD systems is difficult since the training and feedback overhead is proportional to the number of antennas at the base station, which is large in a Massive MIMO system. The large training time is also infeasible due to the limited coherence time of the channel.

To deal with the limited downlink channel training interval in a FDD Massive MIMO system, one option is to explore possible underlying channel structure whereby the high dimensional channel vector has a low dimensional representation [3]–[6]. Motivated by the framework of *Compressive Sensing* (CS), if the desired signal (channel response) can be sparsely represented in some basis or dictionary, then the number of measurements (downlink training period) is proportional to the number of nonzero entries and the signal can be robustly recovered using sparse signal recovery algorithms [7], [8]. In the previous works where CS based channel estimation has been applied, the discrete Fourier transform (DFT) matrix has been used to sparsely represent the channel [4]–[6], [9]. The utilization of DFT basis is compatible with theoretical results of signal recovery in compressive sensing [7], and has been proposed as the virtual channel model [10] or angular domain channel representation [11]. However, the DFT

basis is only valid for a uniform linear array (ULA), and can only lead to an approximate sparse representation with limited scattering and sufficiently large number of antennas [9], [10]. For practical channels, the DFT basis will often result in a large number of nonzero entries in the channel representation, which in turn requires a large number of training symbols for reliable channel estimation, thus losing the benefits of CS based channel estimation.

Motivated by the CS framework, this work extends the framework to provide effective solutions to the FDD Massive MIMO channel estimation problem. In order to accurately and sparsely represent the channel, a *dictionary learning based channel model* (DLCM) is proposed, where a learned *overcomplete* dictionary, rather than a predefined basis or dictionary, is used to represent the channel. The dictionary, due to the learning process, is able to adapt to the cell characteristics as well as insure a sparse representation of the channel. Since no structural constraints are placed on the dictionary, the approach is applicable to an arbitrary array geometry and also does not need accurate array calibration. Learning the basis or dictionary to sparsely represent a signal has been subject of much work and many algorithms have been developed to effectively solve the learning problem [12]–[16]. It has been applied widely in areas of image denoising [17], classification [18], feature extraction [19] and others. Observing the similarity between learning a dictionary for a specific image class and learning a dictionary for a specific cell, we build and expand the framework of dictionary learning to seek the most accurate and efficient representation of the channels in a particular cell. Once the channel has the sparse representation, the compressed channel estimation framework can be utilized to reduce the amount of downlink training.

In Massive MIMO systems, as mentioned before, uplink channel estimation is a relatively easy task. However, in a FDD system since uplink channel and downlink channel are in different frequency bands, channel reciprocity is no longer valid and so one cannot directly use uplink channel for downlink beamforming. However, the propagation environment is the same for the uplink and downlink transmissions. As a result, when the duplex distance is not large, it is reasonable to assume that the number of significant multipaths, the path delays, and path angles are the same for both links [20]–[25]. Because the number and location of the scattering clusters are closely related to the number and location of nonzero entries in the sparse representation, it motivates the use of a jointly sparse dictionary learning approach for the uplink and downlink channel. This enables the use of information from the more easily obtained uplink training to help downlink channel estimation. Leveraging uplink channel information for downlink use in

FDD systems has been proposed in previous work, for example using uplink signals to compute direction of arrival (DOA) and construct downlink channel response [21], [25], or utilizing uplink channel covariance matrix to estimate downlink channel covariance matrix [20], [23], [26], [27]. In [28], a new array configuration is designed such that uplink and downlink second order statistics are the same. In this work, the concept of jointly sparse dictionary learning is proposed as an abstract model for capturing and utilizing the reciprocity that resides in the uplink/downlink channels.

In summary, to address the challenge of downlink channel estimation in a massive MIMO FDD system, we develop and apply the framework of sparse and jointly sparse dictionary learning based approaches. The experimental results show that utilizing the proposed channel models for compressed channel estimation results in significantly reducing the downlink training overhead. The learning from channel measurements also offers the additional benefit of robustness to uncertainties that may arise from numerous factors. The array manifold uncertainty is studied in this work. Preliminary versions of this work have appeared in [29], [30],

A. Contributions

The contributions of this paper are as follows:

- We propose an overcomplete dictionary learning based approach for channel modeling where the spatial Massive MIMO channel can be sparsely represented. Such a learned model has the potential to achieve a more accurate and efficient channel representation by virtue of the dictionary being able to adapt to the channel properties of a cell. This in turn improves the compressed channel estimation and reduces the downlink training overhead. To the best of our knowledge, our work is the first to utilize the dictionary learning framework to model the spatial Massive MIMO channel for FDD systems.
- Observing the reciprocity in the angular domain of FDD uplink/downlink channels, a general framework of joint dictionary learning based channel modeling is proposed to exploiting the same sparsity structure between the uplink and downlink representation. To the best of our knowledge, our work is the first to explore jointly sparse dictionary representations as an abstract model for the downlink and uplink channel reciprocity. Our work also develops appropriate channel estimation algorithms utilizing this joint structure to improve the performance of downlink channel estimation.

B. Notations and Organization

Notations used in this paper are as follows. Upper (lower) bold face letters are used throughout to denote matrices (column vectors). $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^\dagger$ denotes the transpose, Hermitian transpose, and the Moore-Penrose pseudo-inverse. $\mathbf{A}_{\cdot j}$ represents the j -th column of \mathbf{A} . For a vector \mathbf{x} , $\text{diag}(\mathbf{x})$ is a diagonal matrix with entries of \mathbf{x} along its diagonal. $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ denotes the l_1 and l_2 norm. $\|\mathbf{x}\|_0$ represents the number of nonzero entries in \mathbf{x} and is referred to as the l_0 norm. $\text{supp}(\mathbf{x})$ denotes the set of indices such that the corresponding entries of \mathbf{x} are nonzero.

The rest of the paper is organized as follows. Section II describes the conventional channel estimation method and the existing work of compressive sensing based channel estimation. Section III introduces the dictionary learning based channel model and uses it to estimate the downlink channel. In Section IV, we motivate the abstract reciprocity between the uplink and downlink sparse representation, and propose joint dictionary learning based channel model and compressed channel estimation. We experimentally evaluate and compare the performance of different algorithms in Section V. The paper is concluded in Section VI.

II. COMPRESSED CHANNEL ESTIMATION

A. Conventional Channel Estimation

Consider a single cell Massive MIMO system operated in FDD mode. Assume the base station is equipped with N antennas and the user terminal has a single antenna [3], [31]. We consider transmissions over a narrowband block flat-fading channel. For the downlink channel estimation in FDD system, the base station transmits training pilots. Users estimate the channel and feed back the channel state information to the base station. The received signal \mathbf{y}^d at the user terminal is given as

$$\mathbf{y}^d = \sqrt{\rho^d} \mathbf{A} \mathbf{h}^d + \mathbf{n}^d \quad (1)$$

where $\mathbf{A} \in \mathbb{C}^{T^d \times N}$ is the downlink pilots transmitted during the training period of T^d symbols, $\mathbf{h}^d \in \mathbb{C}^{N \times 1}$ denotes the downlink channel response vector, ρ^d denotes the received power, $\mathbf{n}^d \in \mathbb{C}^{T^d \times 1}$ is the received noise vector such that $\mathbf{n}^d \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$. Using conventional channel estimation technique such as *Least Square* (LS) channel estimation [32], the estimated channel is given by

$$\hat{\mathbf{h}}_{LS}^d = (\sqrt{\rho^d} \mathbf{A})^\dagger \mathbf{y}^d \quad (2)$$

where \mathbf{A}^\dagger is the Moore-Penrose pseudoinverse. Robust recovery of \mathbf{h}^d by LS channel estimation requires $T^d \geq N$, which means the number of training period needs to be larger than the number of antennas. In a Massive MIMO system N is very large making this infeasible. Moreover, user needs to feed back channel information to the base station, which also requires feedback resources proportional to channel dimension N . The finite channel coherence time further exacerbates the situation.

On the other hand, uplink channel estimation is relatively easy in a Massive MIMO system. With the same assumption of N antennas at the base station and a single antenna at the user terminal, the uplink training can be written as

$$\mathbf{Y}^u = \sqrt{\rho^u} \mathbf{h}^u \mathbf{a} + \mathbf{N}^u \quad (3)$$

where $\mathbf{h}^u \in \mathbb{C}^{N \times 1}$ is the uplink channel, $\mathbf{a} \in \mathbb{C}^{1 \times T^u}$ denotes the uplink pilots during uplink training period T^u , ρ^u denotes the uplink received power, $\mathbf{Y}^u \in \mathbb{C}^{N \times T^u}$ denotes the received signal at the base station and $\mathbf{N}^u \in \mathbb{C}^{N \times T^u}$ is the received noise with each element to be i.i.d Gaussian with mean 0 and variance σ^2 . Using traditional least square channel estimation, we have

$$\hat{\mathbf{h}}_{LS}^u = \mathbf{Y}^u (\sqrt{\rho^u} \mathbf{a})^\dagger. \quad (4)$$

Since \mathbf{a} is a row vector, there is no restriction on the number of uplink training period T^u for LS channel estimation, showing the relative simplicity of uplink channel estimation compared to downlink channel estimation. However, unlike the TDD system where the uplink channel can be directly applied for the downlink beamforming, due to the channel reciprocity [1], [2], in FDD system such reciprocity is missing since the uplink and downlink are operated in different frequency bands. Consequently, there is a need for effective solutions to estimate the downlink channel in FDD Massive MIMO systems.

B. Compressed Channel Estimation

In order to robustly estimate downlink channel with limited training overhead, compressive sensing based channel estimation, namely compressed channel estimation, has been proposed in previous works [4]–[6], [33], [34]. In the compressive sensing framework, methods to measure a high dimensional signal, e.g. \mathbf{h}^d , have been proposed with much smaller measurements, meaning $T^d < N$, provided the original signal can be sparsely represented in some orthonormal basis

[7]. Assume there exists an orthonormal sparsifying basis $\mathbf{D}^d \in \mathbb{C}^{N \times N}$ such that $\mathbf{h}^d = \mathbf{D}^d \boldsymbol{\beta}^d$, where the representation vector $\boldsymbol{\beta}^d \in \mathbb{C}^{N \times 1}$ is sparse, i.e. $\|\boldsymbol{\beta}^d\|_0 = s \ll N$ ¹. Then the downlink channel estimation can be written as

$$\mathbf{y}^d = \sqrt{\rho^d} \mathbf{A} \mathbf{h}^d + \mathbf{n}^d = \sqrt{\rho^d} \mathbf{A} \mathbf{D}^d \boldsymbol{\beta}^d + \mathbf{n}^d. \quad (5)$$

Given $\mathbf{y}^d, \rho^d, \mathbf{A}$ and \mathbf{D}^d , if we are able to solve for $\boldsymbol{\beta}^d$, then the channel estimate is obtained as $\hat{\mathbf{h}}^d = \mathbf{D}^d \boldsymbol{\beta}^d$. However, (5) is an underdetermined system if we aim to use a small number of training samples $T^d < N$. The system will in general have an infinite number of solutions for $\boldsymbol{\beta}^d$ and the sparsity assumption provides a mechanism to regularize the problem. Consider the minimum sparsity assumption that $\|\boldsymbol{\beta}^d\|_0 = s \ll N$ and assume $\|\mathbf{n}^d\|_2 \leq \epsilon$, then the problem reduces to:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^d &= \arg \min_{\boldsymbol{\beta}^d} \|\boldsymbol{\beta}^d\|_0 \quad \text{subject to} \quad \|\mathbf{y}^d - \sqrt{\rho^d} \mathbf{A} \mathbf{D}^d \boldsymbol{\beta}^d\|_2 \leq \epsilon, \\ \hat{\mathbf{h}}_{CS}^d &= \mathbf{D}^d \hat{\boldsymbol{\beta}}^d. \end{aligned} \quad (6)$$

Notice that the optimization formula in (6) is non-convex, and a number of suboptimal but effective algorithms have been proposed to solve the problem [35]–[39]. One of the most widely used framework is to relax the l_0 norm $\|\boldsymbol{\beta}^d\|_0$ to the l_1 norm $\|\boldsymbol{\beta}^d\|_1$. It has been shown that under certain conditions on $\mathbf{A} \mathbf{D}^d$, based on the ℓ_1 norm criteria a solution of $\boldsymbol{\beta}^d$ with bounded error can be obtained with $T^d \geq c \cdot s \log(N/s)$, where c is some constant [7]. This provides us guidance on how much the training overhead can be reduced if the channel \mathbf{h}^d has the property of sparsity. Instead of using a training period proportional to the channel dimension N , which is large in Massive MIMO system, we can compute a good channel estimate with training period proportional to sparsity level s , which is assumed to be much less than N . This makes downlink channel estimation feasible in a limited training period.

C. Sparsifying Basis/Dictionary

The CS channel estimation procedure shows promise and indicates that a channel representation that is sparse, i.e. \mathbf{h}^d satisfies $\mathbf{h}^d = \mathbf{D}^d \boldsymbol{\beta}^d, \|\boldsymbol{\beta}^d\|_0 \ll N$, is essential to the successful implementation of compressed channel estimation. In this regards, there are two important

¹The zero norm $\|\boldsymbol{\beta}\|_0$ counts the number of nonzero entries in $\boldsymbol{\beta}$. But it is not technically a norm because it does not satisfy the required axioms.

questions: first, does the underlying physical propagation environment support a sparse representation? Second, can we find a good sparsifying basis \mathbf{D}^d such that the mathematical model $\mathbf{h}^d = \mathbf{D}^d \boldsymbol{\beta}^d$, $\|\boldsymbol{\beta}^d\|_0 \ll N$ is valid?

First we discuss some physical channel attributes, somewhat simplified, that support the sparse representation. We adopt a simplified spatial channel model [9], [10], [40] which captures the physical propagation structure of downlink transmission as

$$\mathbf{h}^d = \sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \alpha_{il}^d \mathbf{a}^d(\theta_{il}) \quad (7)$$

where N_c is the number of scattering clusters, each of which contains N_s propagation subpaths. α_{il}^d is the complex gain of the l -th subpath in the i -th scattering cluster. θ_{il} denotes the angle of departure (AOD) for downlink transmission. The vector $\mathbf{a}^d(\theta_{il})$ represents the normalized array response at angle θ_{il} , and for a uniform linear array we have

$$\mathbf{a}^d(\theta_{il}) = \frac{1}{\sqrt{N}} [1, e^{j2\pi \frac{d}{\lambda^d} \sin(\theta_{il})}, \dots, e^{j2\pi \frac{d}{\lambda^d} \sin(\theta_{il}) \cdot (N-1)}]^T \quad (8)$$

where d is the antenna spacing and λ^d is the wavelength of propagation. Unless otherwise specified, the critical spacing $d = \lambda^d/2$ is assumed. In order to model the scattering clusters, we consider the principles of *Geometry-Based Stochastic Channel Model* (GSCM) [41], as illustrated in Fig. 1. For a specific cell, the locations of the dominant scattering clusters are determined by cell specific attributes such as the buildings, terrain, etc and are common to all the users irrespective of user position. We assume such scattering clusters are far away from the base station, so the subpaths associated with a specific scattering cluster will be concentrated in a small range around the line of sight (LOS) direction between the base station and the scattering cluster, i.e. having a small angular spread (AS). While modeling the scattering effects which are user-location dependent, for example the ground reflection close to the user, or some moving physical scatterers near the user, we assume the user terminal is far away from the base station, so subpaths associated with the user-location dependent scattering cluster also have small angular spread. For each link between the base station and the user, the number of scattering clusters that contributes to the channel responses are limited, i.e. N_c is small. Notice that for users at different locations, if they can see the same scattering cluster then their channel responses will contain subpaths with similar AOA/AOD which all concentrate around the LOS direction between the base station and that scattering cluster, a phenomenon known as "shared scatterers"

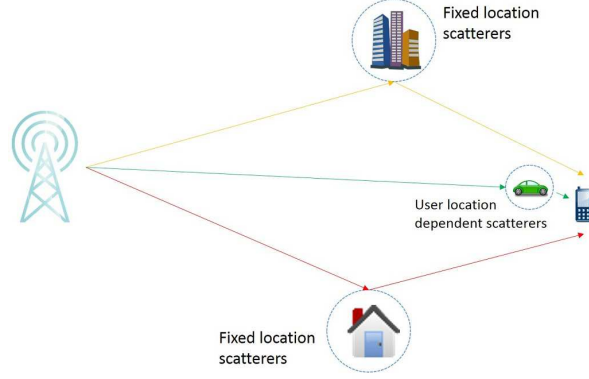


Fig. 1. Illustration of signal propagation in a typical cell

or "joint scatterers" [4], [41], [42]. All these considerations support the idea of a low dimensional representation for the large Massive MIMO channel.

For a uniform linear array, a reasonable basis is the DFT basis \mathbf{F} , which transforms spatial channel response into the angular domain [10], [11], and has the following form:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}(-\frac{1}{2}) & \mathbf{f}(-\frac{1}{2} + \frac{1}{N}) & \dots & \mathbf{f}(\frac{1}{2} - \frac{1}{N}) \end{bmatrix} \in \mathbb{C}^{N \times N}, \quad (9)$$

$$\mathbf{f}(\psi) = \frac{1}{\sqrt{N}} [1, e^{j2\pi\psi}, \dots, e^{j2\pi\psi \cdot (N-1)}]^T.$$

The column of the basis $\mathbf{f}(\psi)$ has the same structure as the array response $\mathbf{a}^d(\theta)$ in (8), and ψ is related to θ through $\psi = d\sin(\theta)/\lambda^d$. Due to the large antenna array at the base station, it has finer resolvability of spatial signals and much lower sidelobes and leakage [43], which again supports the sparse assumption. Since the DFT basis is an orthonormal basis, it is compatible with the assumptions made in deriving the theoretical results concerning the efficacy of the l_1 framework [7]. In many previous works applying compressed channel estimation, DFT basis has been used as the sparsifying basis [4]–[6], [33], [34].

However, in practice signals come from arbitrary directions, so $\psi = d\sin(\theta)/\lambda^d$ rarely resides on the DFT bins $\{-\frac{1}{2}, -\frac{1}{2} + \frac{1}{N}, \dots, \frac{1}{2} - \frac{1}{N}\}$. Even the leakage becomes smaller, it still exists and causes the nearby elements corresponding to the incoming signal to be nonzero. The multiple subpaths in the scattering cluster further exacerbate the situation. When we use the representation $\mathbf{h}^d = \mathbf{F}\boldsymbol{\beta}^d$, there could be a lot of nonzero elements in $\boldsymbol{\beta}^d$, thus requiring large training period T^d to robustly estimate it. To achieve a better sparse representation, one immediate extension of

DFT basis is the overcomplete DFT dictionary $\tilde{\mathbf{F}}$, which has the form

$$\tilde{\mathbf{F}} = \begin{bmatrix} \mathbf{f}(-\frac{1}{2}) & \mathbf{f}(-\frac{1}{2} + \frac{1}{M}) & \dots & \mathbf{f}(\frac{1}{2} - \frac{1}{M}) \end{bmatrix} \in \mathbb{C}^{N \times M}. \quad (10)$$

The columns of $\tilde{\mathbf{F}}$ has the same structure $\mathbf{f}(\psi)$, but the angular domain is sampled (in the sense of ψ) more finely, i.e. $M > N$. The overcomplete dictionary introduces redundancy to the basis, which improves the flexibility of representing the signal as well as the capability of inducing sparsity. Due to non-orthogonality of the overcomplete dictionary, we can no longer rely on the results developed within the traditional compressive sensing framework. Interested readers are referred to [44], [45] for some more analysis considering overcomplete sparsifying dictionary. However, in practice, we can still use the algorithms developed to solve the optimization problem and achieve good performance as shown through extensive numerical studies along with some theoretical evidence [35]–[39]. In the simulation results of Section V, we will demonstrate the advantage of overcomplete DFT dictionary over DFT basis for both sparse representation and sparse recovery.

III. DICTIONARY LEARNING BASED CHANNEL MODEL

A. Motivation

In the previous section, for a ULA we introduced the overcomplete DFT dictionary $\tilde{\mathbf{F}}$ as a simple alternative to the orthonormal DFT basis² \mathbf{F} . Both $\tilde{\mathbf{F}}$ and \mathbf{F} are predefined, and independent of the specific cell properties. This independence makes generation of basis/dictionary very easy, but also loses the ability to more effectively represent the channel by exploring cell specific characteristics. For example both $\tilde{\mathbf{F}}$ and \mathbf{F} uniformly sample the ψ domain, but for a specific cell it is possible that no signals may be received from some directions, then the columns in $\tilde{\mathbf{F}}$ and \mathbf{F} corresponding to those directions will never be used. On the other hand, for directions corresponding to locations of scattering clusters, finer angular sampling can lead to a reduced leakage. Furthermore, predefined basis/dictionary assumes ideal channel response, e.g. far-field plane wave assumption [43], equal antenna gain and antenna spacing, etc, which is not robust to any propagation model mismatch or antenna array uncertainty.

In this paper, we propose a dictionary learning based channel model (DLCM) which learns an overcomplete dictionary. The dictionary learning process adapts the channel model to the

²Similar dictionary constructions are possible for arrays with arbitrary geometry as long as the array manifold is known

channel measurements collected in a cell, which contain the specific cell characteristics³. The sparse representation is also encouraged during the learning process, making compressed channel estimation feasible with reduced training pilots. The learned overcomplete dictionary has the potential to unearth underlying low dimensionality through the learning process, and is robust to any antenna array uncertainties and nonideal propagation schemes. When the knowledge of the underlying physical generation scheme of the channel is imperfect or even incorrect, e.g. antenna gains and locations are different from the nominal values, or there exist near-field scattering clusters, etc, the predefined basis/dictionary is no longer accurate and may cause severe performance degradation. However, the learned dictionary does not have any predefined structural constraints, and is able to tune its own structure to adapt to the channel measurements, thus leading to a robust channel representation. The insight behind dictionary learning is that the high dimensional data (channel response in our case) usually has some structure correlated in some dimensions, and the true degrees of freedom that generate the data is usually small. So by learning from large amount of data, we are able to recover useful underlying structures or models, thus making representation of the data more efficient for the desired application. In our situation, one could view this as big data analytics applied to the physical layer.

B. Dictionary Learning

From now, we denote $\mathbf{D}^d \in \mathbb{C}^{N \times M}$ as the learned dictionary from downlink channel measurements, and to benefit from the flexibility of overcompleteness, we let $N < M$. Assuming we collect L channel measurements as the training samples in a specific cell, the goal is to learn \mathbf{D}^d such that for all the downlink channel responses $\mathbf{h}_i^d, i = 1, \dots, L$, they can be approximated as $\mathbf{h}_i^d \approx \mathbf{D}^d \boldsymbol{\beta}_i^d, \boldsymbol{\beta}_i^d \in \mathbb{C}^{M \times 1}$. The algorithms should be able to address model fitting $\|\mathbf{h}_i^d - \mathbf{D}^d \boldsymbol{\beta}_i^d\|_2$ (accuracy), and encourage small $\|\boldsymbol{\beta}_i^d\|_0$ (efficiency) for sparse representation. If we constrain the

³The channel measurements describe the effect of scattering clusters on the transmitted electromagnetic waves and antenna array. So the underlying structure of channel measurements collected in a cell can reflect the cell specific properties regarding to both scattering clusters and the antenna array.

model mismatch error, then the dictionary learning can be formulated as

$$\begin{aligned} \min_{\substack{\mathbf{D}^d \in \mathcal{C} \\ \beta_1^d, \dots, \beta_L^d}} \quad & \frac{1}{L} \sum_{i=1}^L \|\beta_i^d\|_0 \\ \text{subject to} \quad & \|\mathbf{h}_i^d - \mathbf{D}^d \beta_i^d\|_2 \leq \eta, \forall i \end{aligned} \quad (11)$$

where the constraint set \mathcal{C} is defined as

$$\mathcal{C} = \{\mathbf{D} \in \mathbb{C}^{N \times M}, \text{ s.t. } \|\mathbf{D}_{\cdot j}\|_2 \leq 1, \forall j = 1, \dots, M\} \quad (12)$$

in order to prevent the ambiguity between \mathbf{D}^d and β^d . The solved \mathbf{D}^d in (11) leads to the sparsest representation in the sense of all channel measurements, given the model mismatch tolerance η .

Two similar formulations could alternatively be used. If we know beforehand or want to constrain the sparsity level of each coefficient β_i^d , then we solve:

$$\begin{aligned} \min_{\substack{\mathbf{D}^d \in \mathcal{C} \\ \beta_1^d, \dots, \beta_L^d}} \quad & \frac{1}{L} \sum_{i=1}^L \frac{1}{2} \|\mathbf{h}_i^d - \mathbf{D}^d \beta_i^d\|_2^2 \\ \text{subject to} \quad & \|\beta_i^d\|_0 \leq T_0, \forall i \end{aligned} \quad (13)$$

where T_0 constrains the number of non-zero elements in each β_i^d . In other words, we expect every channel measurement can be represented using T_0 atoms from the learned dictionary, and we solve the dictionary that minimize model mismatch using channel response samples. If we do not have any explicit constraints on model fitting error or sparsity level, we can formulate the dictionary learning process in general form as

$$\min_{\substack{\mathbf{D}^d \in \mathcal{C} \\ \beta_1^d, \dots, \beta_L^d}} \quad \frac{1}{L} \sum_{i=1}^L \frac{1}{2} \|\mathbf{h}_i^d - \mathbf{D}^d \beta_i^d\|_2^2 + \lambda \|\beta_i^d\|_0 \quad (14)$$

where λ is the parameter that trades off the data fitting and sparsity.

To solve the dictionary learning problem (11), (13) and (14), block coordinate descent framework has been applied where each iteration includes alternatively minimizing with respect to either \mathbf{D}^d or $\beta_i^d, \forall i$, while keeping the other fixed. When \mathbf{D}^d is fixed, optimizing $\beta_i^d, \forall i$ is a sparse signal recovery problem [35]–[39]. When we fix $\beta_i^d, \forall i$ and solve for \mathbf{D}^d , many dictionary learning algorithm can be applied [12]–[15], [18]. The convergence of the iteration depends on the specific sparse recovery algorithm and dictionary update algorithm, and to the best of our

knowledge, no general guarantees have been provided. Interested readers are referred to [14], [18], [46], [47] for some discussion about the convergence under specific assumptions. For the purpose of this paper, in Section V, we show experimentally that starting from a reasonable dictionary, e.g. an overcomplete DFT dictionary, the learning algorithm will lead to a dictionary that improves the performance in terms of both sparse representation and sparse recovery.

C. Compressed Channel Estimation using DLCDM

The algorithm of compressed channel estimation using dictionary learning based channel model is summarized in Algorithm 1. It includes a dictionary learning phase, which is usually off line, and a compressed channel estimation algorithm that utilizes the learned dictionary. We use (11) as the formulation for dictionary learning, and apply l_1 relaxation of $\|\beta^d\|_0$ to $\|\beta^d\|_1$ in both dictionary learning and compressed channel estimation. To learn a comprehensive dictionary for users in the whole cell, we need to collect downlink channel measurements from all locations in a specific cell, i.e. *cell specific samples*, and the learned dictionary is used only for this cell. Such channel measurements collection is offline, and assumed to be done at the cell deployment stage. The learned dictionary is stored in the BS for use during downlink training. At this stage there is not much concern about reducing training overhead and one would like to collect channel measurements as accurately and as extensively as possible.

The compressed channel estimation is performed during the training phase. In compressed channel estimation, users feed back the measurements, made during the training phase, to the base station and the channel is estimated at the base station using the learned dictionary [4], [48]. It is different from the conventional channel estimation where users estimate the channel and feed back the channel state information to the base station. The scheme in Algorithm 1 has several advantages: first the sparse recovery algorithms (channel estimation) can be complex and so it is preferably done at the base station thus saving energy for user terminals. Secondly, the dictionary is learned and stored at the base station, because the alternative of making it available to all users involves significant overhead in storage at the user equipment and also conveyance of dictionary. Since the received symbol \mathbf{y}^d has dimension T^d which is much less than the channel dimension N in Massive MIMO system, the scheme in Algorithm 1 also reduces feedback overhead which is now only proportional to the channel sparsity level. In this work, we assume perfect uplink feedback for simplicity.

Algorithm 1 Compressed Channel Estimation Using DLCM

Dictionary Learning Phase

Input: model mismatch constraint η .

Output: learned dictionary D^d .

- Collect downlink channel measurements $\mathbf{h}_i^d, i = 1, \dots, L$ in different locations of the cell.
- Solve the dictionary learning problem

$$\min_{\substack{D^d \in \mathcal{C} \\ \beta_1^d, \dots, \beta_L^d}} \frac{1}{L} \sum_{i=1}^L \|\beta_i^d\|_1 \text{ s.t. } \|\mathbf{h}_i^d - D^d \beta_i^d\|_2 \leq \eta, \forall i.$$

Compressed Channel Estimation Phase

Input: downlink pilots \mathbf{A} , downlink dictionary D^d , downlink received power ρ^d , error constraint ϵ

Output: downlink channel estimation $\hat{\mathbf{h}}^d$

- Base station transmits downlink pilots \mathbf{A} .
- User receives $\mathbf{y}^d = \sqrt{\rho^d} \mathbf{A} \mathbf{h}^d + \mathbf{n}^d$.
- User feeds back \mathbf{y}^d .
- Base station solves sparse recovery

$$\min_{\beta^d} \|\beta^d\|_1 \text{ s.t. } \|\mathbf{y}^d - \sqrt{\rho^d} \mathbf{A} D^d \beta^d\|_2 \leq \epsilon$$

- Base station estimates downlink channel as $\hat{\mathbf{h}}^d = D^d \beta^d$
-

IV. UPLINK/DOWNLINK JOINT DICTIONARY LEARNING BASED CHANNEL MODEL

A. Motivation

In compressed channel estimation, larger training period T^d leads to better recovery performance since we collect more information about the downlink channel \mathbf{h}^d . However, larger T^d also requires more downlink resources for channel estimation and leaves less time for actual data transmission. This motivates our search for alternative information sources that can facilitate downlink channel estimation. For this we draw inspiration from TDD systems, where through channel reciprocity the uplink channel estimate provides downlink channel information [1], [2]. In FDD system we do not have such channel reciprocity because uplink and downlink

transmission are operated in different frequency bands. However, if the duplex distance is not large, i.e. the frequency difference between uplink and downlink is not large, a looser and more abstract form of reciprocity is possible and appropriate. For instance, it is reasonable to assume the AOA of signals in the uplink transmission is the same as the AOD of signals in the downlink transmission [20]–[25], in other words directions of signal paths are invariant to carrier frequency shift. In [24], congruence of the directional properties of uplink and downlink channel is observed experimentally, where the dominant uplink/downlink directions of arrival (DOA) show only a minor deviation, and the uplink/downlink azimuth power spectrums (APS) have a high correlation. Similar to the spatial channel model for the downlink channel response as in (7), we have the spatial channel model for both uplink and downlink as following:

$$\begin{aligned} \mathbf{h}^u &= \sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \alpha_{il}^u \mathbf{a}^u(\theta_{il}), & \mathbf{h}^d &= \sum_{i=1}^{N_c} \sum_{l=1}^{N_s} \alpha_{il}^d \mathbf{a}^d(\theta_{il}), \\ \mathbf{a}^u(\theta_{il}) &= \frac{1}{\sqrt{N}} [1, e^{j2\pi \frac{d}{\lambda^u} \sin(\theta_{il})}, \dots, e^{j2\pi \frac{d}{\lambda^u} \sin(\theta_{il}) \cdot (N-1)}]^T, \\ \mathbf{a}^d(\theta_{il}) &= \frac{1}{\sqrt{N}} [1, e^{j2\pi \frac{d}{\lambda^d} \sin(\theta_{il})}, \dots, e^{j2\pi \frac{d}{\lambda^d} \sin(\theta_{il}) \cdot (N-1)}]^T \end{aligned} \quad (15)$$

where $\alpha_{il}^u, \alpha_{il}^d$ denote the uplink/downlink complex path gain and λ^u, λ^d are the uplink/downlink wavelength. Due to frequency separation of uplink/downlink, α_{il}^u is different and uncorrelated from α_{il}^d . But both link share the same N_c, N_s and θ_{il} . So when we treat \mathbf{h}^u and \mathbf{h}^d as a whole, they appear to be uncorrelated. However, if we are able to resolve them finely in the angular domain, which actually can be achieved by large antenna array, they will show the common spatial structure which can be treated as the reciprocity in the angular domain.

As demonstrated in the previous sections, the AOA/AOD in the angular domain are closely related to the locations of nonzero entries in channel sparse coefficients. Motivated by the DLCM for downlink channel responses, we consider a *joint dictionary learning based channel model* (JDLCM) to find sparse representations of both uplink and downlink channel responses. Similar to the downlink DLCM where $\mathbf{h}^d \approx \mathbf{D}^d \boldsymbol{\beta}^d$, we assume there exists an uplink dictionary \mathbf{D}^u such that $\mathbf{h}^u \approx \mathbf{D}^u \boldsymbol{\beta}^u$. Then the same AOA/AOD translates to the same locations of nonzero entries in $\boldsymbol{\beta}^u$ and $\boldsymbol{\beta}^d$, i.e. $\text{supp}(\boldsymbol{\beta}^u) = \text{supp}(\boldsymbol{\beta}^d)$. Consequently, if we know \mathbf{h}^u , and utilize for the downlink channel estimation the common support information, i.e. $\text{supp}(\boldsymbol{\beta}^u) = \text{supp}(\boldsymbol{\beta}^d)$, we have critical information about \mathbf{h}^d and achieve better downlink channel estimates without increasing the downlink training overhead.

B. Joint Dictionary Learning Based Channel Model

Similar to the DLCM for the downlink, we aim to learn an uplink dictionary \mathbf{D}^u from collection of uplink channel measurements. Furthermore, \mathbf{D}^u and \mathbf{D}^d are learned jointly with constraint $\text{supp}(\beta^u) = \text{supp}(\beta^d)$, leading to a joint dictionary learning based channel model (JDLCM). In order to enforce the same support constraint $\text{supp}(\beta^u) = \text{supp}(\beta^d)$, we collect pairs of channel measurements $\{h_i^u, h_i^d\}, i = 1, \dots, L$ all across the cell. Each pair of channel measurements $\{h_i^u, h_i^d\}$ are measured at the same user location, so the assumption of the same AOA/AOD is valid, which implies $\text{supp}(\beta^u) = \text{supp}(\beta^d)$. The joint dictionary learning problem can be formulated as

$$\begin{aligned} \min_{\substack{\mathbf{D}^u \in \mathcal{C}, \beta_1^u, \dots, \beta_L^u \\ \mathbf{D}^d \in \mathcal{C}, \beta_1^d, \dots, \beta_L^d}} \quad & \frac{1}{L} \sum_{i=1}^L \|\beta_i^u\|_0 + \|\beta_i^d\|_0 \\ \text{subject to} \quad & \|\mathbf{h}_i^u - \mathbf{D}^u \beta_i^u\|_2 \leq \eta, \quad \|\mathbf{h}_i^d - \mathbf{D}^d \beta_i^d\|_2 \leq \eta, \\ & \text{supp}(\beta_i^u) = \text{supp}(\beta_i^d), \quad \forall i \end{aligned} \tag{16}$$

which is very similar to the downlink dictionary learning problem as shown in (11), except the constraint $\text{supp}(\beta_i^u) = \text{supp}(\beta_i^d)$. This constraint is important since it builds the connection between the uplink and downlink channel responses, which will be utilized in the channel estimation. Two alternative joint dictionary learning formulation can be extended from (13) and (14) straightforwardly, and we omit them here due to space limitation.

To solve the joint dictionary learning, we minimize (16) iteratively, i.e. we fix $\mathbf{D}^u, \mathbf{D}^d$ and solve for $\beta_i^u, \beta_i^d, \forall i$, and then fix $\beta_i^u, \beta_i^d, \forall i$ and solve for $\mathbf{D}^u, \mathbf{D}^d$. Notice that when $\beta_i^u, \beta_i^d, \forall i$ are fixed, the solution of \mathbf{D}^u and \mathbf{D}^d are decoupled, and can be optimized independently using any of dictionary learning algorithms [12]–[15], [18]. β_i^u, β_i^d are coupled through the constraint $\text{supp}(\beta_i^u) = \text{supp}(\beta_i^d)$, and need to be solved jointly. Joint sparse recovery using different dictionary $\mathbf{D}^u, \mathbf{D}^d$ and with structure such as $\text{supp}(\beta_i^u) = \text{supp}(\beta_i^d)$ has been considered in [49] and termed *distributed compressive sensing*. An OMP like algorithm was proposed to solve the joint recovery problem. In [50], similar joint recovery problem is solved using a Bayesian framework and is called *multi-task learning*. In [51], a reweighted l_p norm like algorithm [52] and a sparse Bayesian learning like algorithm [53] has been proposed for joint sparse recovery. It has been shown that joint recovery can lead to more accurate results compared to independent recovery. In this paper, we consider a group l_1 formulation which is similar to the group-lasso

[54] to solve the joint sparse recovery problem. Assume $\mathbf{D}^u, \mathbf{D}^d$ are fixed, then pairs of $\{\beta_i^u, \beta_i^d\}$ are decoupled and each pair can be solved independently. By forming

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}_i^d \\ \mathbf{h}_i^u \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{D}^d & \mathbf{0}_{N \times M} \\ \mathbf{0}_{N \times M} & \mathbf{D}^u \end{bmatrix} \quad (17)$$

the joint sparse recovery of β_i^d, β_i^u can be written as

$$\min_{\beta} \sum_{j=1}^M \|\beta\|_{\mathbf{K}_j} \quad (18)$$

$$\text{subject to } \|\mathbf{h} - \mathbf{G}\beta\|_2 \leq \eta$$

where $\|\beta\|_{\mathbf{K}_j} = (\beta^H \mathbf{K}_j \beta)^{1/2}$, \mathbf{K}_j is the group kernel and $\mathbf{K}_j = \text{diag}([e_j^T e_j^T])$, where $e_j \in \mathbb{R}^{M \times 1}$ is the standard basis with 1 on the j -th location and 0 elsewhere. The solved β is the combined vector of the form $\beta = [(\beta_i^d)^T (\beta_i^u)^T]^T$, where we can extract β_i^d, β_i^u . The cost function in (18) is a l_2/l_1 norm of β similar to l_p/l_1 norm in [52], which encourages all the elements in the group to be zero or nonzero simultaneously, and the number of groups to be as small as possible. By applying this group l_1 framework, we enforce the constraint of $\text{supp}(\beta_i^u) = \text{supp}(\beta_i^d)$ and encourage a sparse representation.

C. Compressed Channel Estimation Using JDLCM

After learning $\mathbf{D}^u, \mathbf{D}^d$, we have the joint uplink/downlink sparse channel representation as $\mathbf{h}^u \approx \mathbf{D}^u \beta^u$ and $\mathbf{h}^d \approx \mathbf{D}^d \beta^d$. The goal is to utilize uplink training to help improving the performance of the downlink channel estimation, by using the constraint $\text{supp}(\beta^u) = \text{supp}(\beta^d)$. Given $\mathbf{Y}^u = \sqrt{\rho^u} \mathbf{h}^u \mathbf{a} + \mathbf{N}^u$, it is easy to suppress the noise by sending the same pilot symbol, e.g. p , in T^u uplink training period, i.e. $\mathbf{a} = p \cdot \mathbf{1}_{1 \times T^u}$, where $\mathbf{1}_{1 \times T^u}$ is a $1 \times T^u$ vector with all 1's. Then denote $\mathbf{y}^u = \frac{1}{T^u} \sum_{i=1}^{T^u} \mathbf{Y}_{\cdot i}^u$ and $\mathbf{n}^u = \frac{1}{T^u} \sum_{i=1}^{T^u} \mathbf{N}_{\cdot i}^u$, we have

$$\mathbf{y}^u = \sqrt{\rho^u} p \mathbf{h}^u + \mathbf{n}^u = \sqrt{\rho^u} p \mathbf{D}^u \beta^u + \mathbf{n}^u. \quad (19)$$

Combined with the downlink training in (5), the compressed channel estimation using JDLCM can be formulated as

$$\begin{aligned} \{\hat{\beta}^u, \hat{\beta}^d\} &= \arg \min_{\beta^u, \beta^d} \|\beta^u\|_0 + \|\beta^d\|_0 \\ \text{subject to } &\|\mathbf{y}^u - \sqrt{\rho^u} p \mathbf{D}^u \beta^u\|_2 \leq \epsilon^u, \\ &\|\mathbf{y}^d - \sqrt{\rho^d} \mathbf{A} \mathbf{D}^d \beta^d\|_2 \leq \epsilon^d, \\ &\text{supp}(\beta^u) = \text{supp}(\beta^d). \end{aligned} \quad (20)$$

And the uplink and downlink channel can be estimated as $\hat{\mathbf{h}}^u = \mathbf{D}^u \hat{\beta}^u$ and $\hat{\mathbf{h}}^d = \mathbf{D}^d \hat{\beta}^d$. Again, we face the same joint sparse recovery problem with structure constraint $\text{supp}(\beta^u) = \text{supp}(\beta^d)$ as in the joint dictionary learning problem. We utilize the same group l_1 algorithm as in (18) as follow

$$\begin{aligned} \min_{\beta} &\sum_{j=1}^M \|\beta\|_{\mathbf{K}_j} \\ \text{subject to } &\|\mathbf{y} - \mathbf{G}\beta\|_2 \leq \epsilon \end{aligned} \quad (21)$$

where now we have

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}^d \\ \mathbf{y}^u \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \sqrt{\rho^d} \mathbf{A} \mathbf{D}^d & \mathbf{0}_{T \times M} \\ \mathbf{0}_{N \times M} & \sqrt{\rho^u} p \mathbf{D}^u \end{bmatrix} \quad (22)$$

and the same definition of \mathbf{K}_j . Also the solved β has the form of $\beta = [(\beta^d)^T (\beta^u)^T]^T$. By joint sparse recovery of β^u, β^d , we are able to achieve improved downlink channel estimates with the help of uplink training. Notice the dimension of \mathbf{y}^u is N while dimension of \mathbf{y}^d is T . In the Massive MIMO system where $N \gg T$, the uplink training actually has larger number of measurements, which is beneficial for the sparse recovery in compressive sensing. We can also improve the signal to noise ratio of the uplink received signal by increasing the uplink training period T^u . Due to the constraint $\text{supp}(\beta^u) = \text{supp}(\beta^d)$, \mathbf{y}^u and \mathbf{y}^d can regularize each other to achieve better recovery performance compared to independent recovery. More importantly, the performance of downlink compressed channel estimation is improved without increasing the downlink training period T^d .

We summarize the algorithm of compressed channel estimation using JDLCM in Algorithm 2. Similar to Algorithm 1, there is a dictionary learning phase and a channel estimation phase. During the dictionary learning phase, a large amount of channel measurements need to be collected as training samples. Each pair of uplink/downlink channel measurements has to be

collected at the same user location, in order to guarantee the same AOA/AOD for the uplink and downlink. This requirement is important since for each pair of $\{\mathbf{h}_i^u, \mathbf{h}_i^d\}$ the learning process has the constraint $\text{supp}(\beta_i^u) = \text{supp}(\beta_i^d)$. The joint dictionary learning is implemented when the cell is installed, and the learned $\mathbf{D}^u, \mathbf{D}^d$ are stored at the base station for joint compressed channel estimation. In the channel estimation phase of each coherence time, the base station transmits downlink pilots while the user transmits uplink pilots, then the user feeds back the received signal. The joint channel estimation is performed at the base station, from which the downlink channel state information is obtained.

V. SIMULATION RESULTS

In this section, we experimentally illustrate the advantage of dictionary learning based modeling for efficiently representing and estimating the channel, and demonstrate the further improved performance by applying joint dictionary learning and compressed channel estimation. We assume the base station has a uniform linear array with $N = 100$ antennas, and the user terminal has 1 antenna, so both $\mathbf{h}^d \in \mathbb{C}^{100 \times 1}$ and $\mathbf{h}^u \in \mathbb{C}^{100 \times 1}$. The channel coefficients are generated following the principles of Geometry-Based Stochastic Channel Model (GSCM) [41] with parameters according to [40]. We assume a urban macrocell with radius of 1200 meters centered at the base station, and the azimuth angle ranges from $-\pi/2$ to $\pi/2$ with respect to the broadside of the base station antenna array. The locations of the fixed scattering clusters are drawn uniformly between the range 300 meters to 800 meters at the beginning of the simulation, and then kept constant. There are total of 7 fixed location scattering clusters. The user location is drawn uniformly between the range 500 meters to 1200 meters. For each channel measurement, it includes 4 scattering clusters, where 3 of them are from the fixed scattering clusters that are closest to the user, while the remaining cluster is at the user location in order to model user-location dependent scatterers. Each cluster contains 20 subpaths concentrated in a 4 degree angular spread. The AOA/AOD values are uniformly random generated within the angular spread and identical between the uplink and downlink, and the random subpath phases between uplink and downlink are uncorrelated [40]. We generate $L = 10000$ downlink channel responses $\mathbf{h}_i^d, i = 1, \dots, L$ for downlink dictionary learning, and $L = 10000$ pairs of uplink/downlink channel responses $\{\mathbf{h}_i^u, \mathbf{h}_i^d\}, i = 1, \dots, L$ for joint dictionary learning. In this

Algorithm 2 Compressed Channel Estimation Using JDLCM

Joint Dictionary Learning Phase

Input: model mismatch constraint η .

Output: learned dictionary D^u, D^d .

- Collect pairs of uplink/downlink channel measurements $\{\mathbf{h}_i^u, \mathbf{h}_i^d\}, i = 1, \dots, L$ in different locations of the cell.
- Solve the dictionary learning problem

$$\begin{aligned} \min_{\substack{D^u \in \mathcal{C}, \beta_1^u, \dots, \beta_L^u \\ D^d \in \mathcal{C}, \beta_1^d, \dots, \beta_L^d}} \quad & \frac{1}{L} \sum_{i=1}^L \|\beta_i^u\|_0 + \|\beta_i^d\|_0 \\ \text{subject to} \quad & \|\mathbf{h}_i^u - D^u \beta_i^u\|_2 \leq \eta, \|\mathbf{h}_i^d - D^d \beta_i^d\|_2 \leq \eta, \\ & \text{supp}(\beta_i^u) = \text{supp}(\beta_i^d), \quad \forall i \end{aligned}$$

where the joint sparse recovery step is solved using (18)

Joint Compressed Channel Estimation Phase

Input: uplink/downlink pilots p and \mathbf{A} , uplink/downlink dictionary D^u and D^d , uplink/downlink received power ρ^u and ρ^d , error constraint ϵ

Output: downlink channel estimation $\hat{\mathbf{h}}^d$

- Base station transmits downlink pilots \mathbf{A} .
- User receives $\mathbf{y}^d = \sqrt{\rho^d} \mathbf{A} \mathbf{h}^d + \mathbf{n}^d$.
- User terminal transmits uplink pilots $\mathbf{a} = p \cdot \mathbf{1}_{1 \times T^u}$.
- Base station receives $\mathbf{Y}^u = \sqrt{\rho^u} \mathbf{h}^u \mathbf{a} + \mathbf{N}^u$.
- Base station calculates $\mathbf{y}^u = \frac{1}{T^u} \sum_{i=1}^{T^u} \mathbf{Y}_i^u$.
- User feeds back \mathbf{y}^d .
- Base station solves joint sparse recovery

$$\min_{\beta} \sum_{j=1}^M \|\beta\|_{\mathbf{K}_j}$$

$$\text{subject to } \|\mathbf{y} - \mathbf{G}\beta\|_2 \leq \epsilon$$

where \mathbf{y} and \mathbf{G} are given in (22).

- Base station estimates downlink channel as $\hat{\mathbf{h}}^d = D^d \beta^d$.
-

paper, we apply K-SVD⁴ [14] algorithm for dictionary learning. For sparse recovery and joint sparse recovery in the dictionary learning and compressed channel estimation, we utilize l_1 and group l_1 framework implemented by SPGL1 toolbox⁵ [55]. There are many other dictionary learning and sparse recovery algorithms, which trade off between accuracy and speed. In this work our main objective is to demonstrate the usefulness and potential of dictionary based channel modeling, and leave the problem of optimal algorithm design/selection for future work.

A. Sparse Representation Using DLCM

The motivation of using DLCM is to find a dictionary which can more efficiently represent the channel response, i.e. the channel response can be represented by a sparse vector with fewer number of nonzero entries. We compare three different sparsifying basis and dictionaries: 1) The $N \times N$ DFT basis \mathbf{F} in (9), which has been used in previous works when compressed channel estimation is applied. 2) The $N \times M$ overcomplete DFT dictionary $\tilde{\mathbf{F}}$ in (10), which is an extension of \mathbf{F} suggested in Section II-C. 3) The $N \times M$ overcomplete learned dictionary \mathbf{D}^d . We set $M = 400$ corresponding to a redundancy factor of 4. We generate $L = 1000$ downlink channel responses $\mathbf{h}_i^d, i = 1, \dots, L$ with user location uniformly drawn in the cell, and calculate sparse coefficient β_i^d of each channel responses \mathbf{h}_i^d using the learned dictionary \mathbf{D}^d as

$$\arg \min_{\beta_i^d} \|\beta_i^d\|_1 \text{ subject to } \|\mathbf{h}_i^d - \mathbf{D}^d \beta_i^d\|_2 \leq \eta \quad (23)$$

and compute $\|\beta_i^d\|_0$ and $\|\beta_i^d\|_1$. Similar calculation is done by changing the \mathbf{D}^d in (23) to \mathbf{F} and $\tilde{\mathbf{F}}$.

We plot the cumulative distribution (CDF) of $\|\beta_i^d\|_0$ and $\|\beta_i^d\|_1$ calculated using \mathbf{D}^d , $\tilde{\mathbf{F}}$ and \mathbf{F} in Fig. 2(a) and Fig. 2(b), with $\eta = 0.01$. The CDF of $\|\beta_i^d\|_0$ shows how many columns in \mathbf{D}^d (or $\tilde{\mathbf{F}}$ and \mathbf{F}) are needed in order to represent the channel within the precision $\|\mathbf{h}_i^d - \mathbf{D}^d \beta_i^d\|_2 \leq \eta$. It is shown that fewer number of columns are required from \mathbf{D}^d compared to $\tilde{\mathbf{F}}$ and \mathbf{F} . For example, 60 columns from \mathbf{D}^d are enough to represent 80% of all channel responses, while it requires 80 columns from $\tilde{\mathbf{F}}$ and almost 100 columns from \mathbf{F} . Since the optimization function in (23) is with respect to the l_1 norm of β_i^d , and l_1 norm is a convex relaxation of l_0 norm

⁴K-SVD toolbox: <http://www.cs.technion.ac.il/~elad/software/>

⁵SPGL1 toolbox: <http://www.cs.ubc.ca/~mpf/spgl1/citing.html>

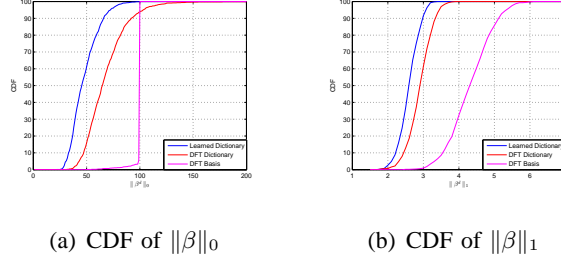


Fig. 2. Cumulative distribution function of $\|\beta\|_0$ and $\|\beta\|_1$ for a base station with 100 antennas. Model mismatch factor $\eta = 0.01$.

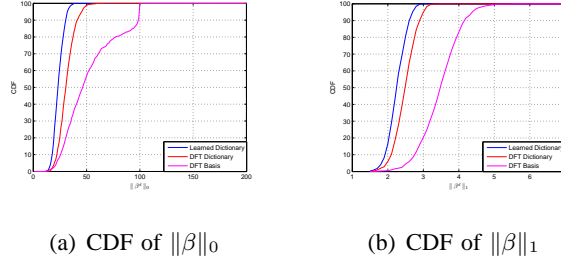


Fig. 3. Cumulative distribution function of $\|\beta\|_0$ and $\|\beta\|_1$ for a base station with 100 antennas. Model mismatch factor $\eta = 0.1$.

which also indicates the sparsity, we plot CDF of $\|\beta_i^d\|_1$ in Fig. 2(b). It indicates that the sparse representation has smaller l_1 norm when \mathbf{D}^d is used to represent the channel response, and both \mathbf{D}^d and $\tilde{\mathbf{F}}$ are much better than \mathbf{F} , which is used in most (if not all) existing works that apply compressed channel estimation.

In Fig. 3, we plot CDF of $\|\beta_i^d\|_0$ and $\|\beta_i^d\|_1$ with $\eta = 0.1$, which means we require a less accurate channel representation. As before, \mathbf{D}^d and $\tilde{\mathbf{F}}$ are much better than \mathbf{F} , while \mathbf{D}^d is better than $\tilde{\mathbf{F}}$. Comparing to Fig. 2, the l_0 norm and the l_1 norm of the sparse coefficient are both largely reduced, due to the reduced accuracy of the channel representation. This is intuitive since we can always drop some entries in β_i^d to reduce $\|\beta_i^d\|_0$ and $\|\beta_i^d\|_1$, while sacrificing the accuracy in the representation of \mathbf{h}_i^d . The results also indicate that we only need small number of columns to coarsely represent the channel. For example, when $\eta = 0.1$ all channel responses can be represented by 40 columns in \mathbf{D}^d . But precisely representing the channel would require many more columns. For example, 40 columns can only represent about 40% of channel responses when $\eta = 0.01$.

B. Compressed Channel Estimation Using DLCM

In this subsection, we compare the performance of compressed channel estimation, when using different sparsifying basis and dictionaries. We generate $L = 1000$ channel responses \mathbf{h}_i^d and calculate the downlink channel training $\mathbf{y}^d = \sqrt{\rho^d} \mathbf{A} \mathbf{h}^d + \mathbf{n}^d$. Element of \mathbf{A} is i.i.d Gaussian random variable with mean 0 and variance 1. $\mathbf{n}^d \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ and we set $\sigma = 1$. The \mathbf{h}^d is normalized such that $\|\mathbf{h}^d\|_2 = 1$, and $\sqrt{\rho^d}$ represents the combined effect of transmitted power and large scale fading of \mathbf{h}^d . The downlink signal to noise ratio is defined by

$$\text{SNR} = \frac{\mathbb{E} \|\sqrt{\rho^d} \mathbf{A} \mathbf{h}^d\|_2^2}{\mathbb{E} \|\mathbf{n}^d\|_2^2} = \rho^d. \quad (24)$$

Given \mathbf{y}^d , \mathbf{A} and \mathbf{D} (or $\tilde{\mathbf{F}}$, \mathbf{F}), we apply the compressed channel estimation using l_1 relaxation as detailed in Algorithm 1, and compare the normalized mean square error (NMSE) as

$$\text{NMSE} = \mathbb{E} \frac{\|\mathbf{h}^d - \hat{\mathbf{h}}^d\|_2^2}{\|\mathbf{h}^d\|_2^2} = \frac{1}{L} \sum_{i=1}^L \|\mathbf{h}_i^d - \hat{\mathbf{h}}_i^d\|_2^2. \quad (25)$$

In Fig. 4, we plot the NMSE of compressed channel estimation using \mathbf{D} , $\tilde{\mathbf{F}}$ and \mathbf{F} , with respect to the number of downlink training symbols T^d . We also include the least square channel estimation when $T^d = 100$. It is shown that even with much less training symbols T^d , the compressed channel estimation technique still achieves much better performance compared to the least square channel estimate, which demonstrates the advantage of compressed channel estimation for Massive MIMO system. Comparing different sparsifying basis/dictionary, Fig. 4 shows that both \mathbf{D} and $\tilde{\mathbf{F}}$ are much better than \mathbf{F} , while \mathbf{D} is even better than $\tilde{\mathbf{F}}$. For example, about a 100 downlink training period is required to achieve $\text{NMSE} = 10^{-2}$ when \mathbf{F} is applied, while only about 50 and 55 training interval is needed when \mathbf{D} and $\tilde{\mathbf{F}}$ are applied. So compressed channel estimation using DLCM saves 5 and 50 training symbols compared to DFT dictionary and DFT basis. There are more advantages in applying \mathbf{D} over $\tilde{\mathbf{F}}$ when we have stricter requirement on NMSE. For example when requiring $\text{NMSE} = 3 \times 10^{-3}$, it requires about 75 downlink training symbols when utilizing \mathbf{D} and 85 training symbols when utilizing $\tilde{\mathbf{F}}$. In summary, compressed channel estimation using DLCM can achieve the same performance as using DFT dictionary and DFT basis with fewer number of downlink training symbols, thus reducing the training overhead.

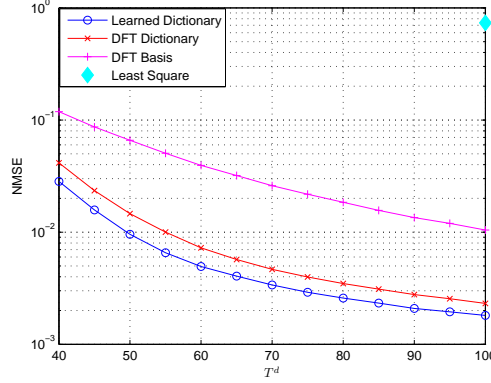


Fig. 4. Normalized mean square error (NMSE) comparison of different sparsifying basis and dictionaries. SNR = 30dB.

C. Dealing with Array Uncertainties Using DLCM

In the previous subsections, although compressed channel estimation using DLCM can achieve the best performance, the DFT dictionary also performs well. The reason is that the DFT dictionary is also a reasonably good sparsifying dictionary which matches the channel measurements. The array response $\mathbf{a}(\theta)$ which generates the channel is the same as the columns of the DFT dictionary $\mathbf{f}(\psi)$ if we let $\psi = \frac{d}{\lambda} \sin(\theta)$. However, in practices there exist some mismatches in the channel model as well as uncertainties in the antenna array. For example there may exist some near-field scattering clusters, whose array response can not be modeled using far-field plane waves. Also due to the calibration error, the spacing between the antennas may not be exactly the same, and the antenna gains of different antennas may also vary. Moreover, such mismatches and antenna uncertainties are usually unknown beforehand and fixed over time, e.g. fixed antenna gain and antenna spacing that are deviating from the nominal values. In this subsection, we compare compressed channel estimation using \mathbf{D} , $\tilde{\mathbf{F}}$ and \mathbf{F} when such unknown fixed uncertainties are present.

Assume the antenna array at the base station has gain and spacing perturbations. We generate the antenna gains like following: out of $N = 100$ antennas, there are 20 antennas having gains $1 + e$ while the other 80 antennas have gain 1. The $e \sim \mathcal{CN}(0, 0.1)$, and if $1 + e > 1.2$ or $1 + e < 0.8$, then the gain is set to be 1.2 or 0.8. Once the gains are generated, they are fixed in the whole simulation. The antenna spacing is generated similarly: out of 99 antenna spacing, there are 20 having values $d = (1 + v)\lambda^d/2$ where $v \sim \mathcal{CN}(0, 0.1)$, if $1 + v > 1.2$ or $1 + v < 0.8$,

then the spacing is set to be $1.2\lambda^d/2$ or $0.8\lambda^d/2$. The rest of antenna spacings have values $d = \lambda^d/2$. After the antenna spacings are generated, they are fixed in the whole simulation. We then perform the same experiments in the previous subsections. In Fig. 5 and Fig. 6, CDF of $\|\beta\|_0$ and $\|\beta\|_1$ are plotted with precision of $\eta = 0.01$ and $\eta = 0.1$. Comparing to Fig. 2 and Fig. 3, it shows that the requirement of $\|\beta\|_0$ and $\|\beta\|_1$ to achieve the same percentage increases significantly for the DFT dictionary and DFT basis when there exist antenna array uncertainties. However, for the learned dictionary, it can achieve nearly the same performance. In Fig. 7, we compare the NMSE of different sparsifying basis and dictionaries when antenna array uncertainties exist. It shows that the performance of DFT dictionary degrades considerably, while the learned dictionary achieves almost the same accuracy as in Fig. 4. The result shows that when there are perturbations in the antenna gains and locations, the DFT dictionary is no longer a good dictionary to sparsely represent the channel responses, while the learned dictionary is still a good one. As a result, the compressed channel estimation using DLCCM can greatly reduce the number of downlink training comparing to DFT dictionary and DFT basis, showing its robustness to the antenna uncertainties.

Notice that in Fig. 5 and Fig. 6 the learned dictionary can achieve very *similar* and *sparse* representation (CDF of $\|\beta\|_0$ and $\|\beta\|_1$) of channel responses as in Fig. 2 and Fig. 3. Due to the same sparsity inducing channel generation scheme, i.e. limited number of scattering clusters N_c , and all N_s subpaths are restricted to a small angular spread, even with antenna array uncertainties the underlying degrees of freedom of the environment are still limited. Also since we apply the same cell layout and generation scheme of channel responses, e.g. a total of 7 fixed location scatterers, each channel response consists 3 fixed location scatterers and 1 user-location dependent scatterer, etc, similar $\|\beta\|_0$ and $\|\beta\|_1$ can be achieved even in the presence of antenna array uncertainties. The results imply that the richness of the channel, or the degree of freedom of the channel, is only determined by the environment between the base station and the user, i.e. cell specific. So to most efficiently and accurately represent the channel, it is useful to learn a cell specific dictionary for each cell.

D. Compressed Channel Estimation Using JDLCM

In this subsection, we compare the compressed channel estimation using JDLCM with the three schemes in the previous subsections that do not utilize the uplink training information.

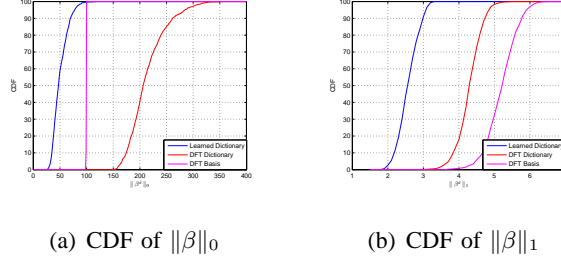


Fig. 5. Cumulative distribution function of $\|\beta\|_0$ and $\|\beta\|_1$ with antenna array uncertainties for a base station with 100 antennas. Model mismatch factor $\eta = 0.01$.

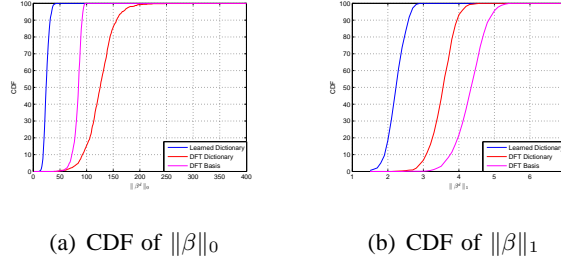


Fig. 6. Cumulative distribution function of $\|\beta\|_0$ and $\|\beta\|_1$ with antenna array uncertainties for a base station with 100 antennas. Model mismatch factor $\eta = 0.1$.

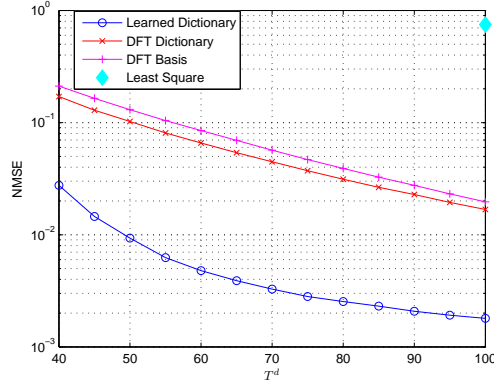


Fig. 7. Normalized mean square error (NMSE) comparison of different sparsifying basis and dictionaries with antenna array uncertainties. SNR = 30dB.

Assume the uplink frequency is 1920 MHz and downlink frequency is 2110 MHz. The antenna spacing $d = \frac{c}{2f_0}$ where c denotes the light speed and $f_0 = 2010$ MHz. We generate $L = 10000$ uplink/downlink channel responses pair $\{\mathbf{h}_i^u, \mathbf{h}_i^d\}$ to perform the joint dictionary learning. In the uplink training $\mathbf{Y}^u = \sqrt{\rho^u} \mathbf{h}^u \mathbf{a} + \mathbf{N}^u$, $\mathbf{a} = p \cdot \mathbf{1}_{1 \times T^u}$ where $p \sim \mathcal{CN}(0, 1)$, element of \mathbf{N}^u are

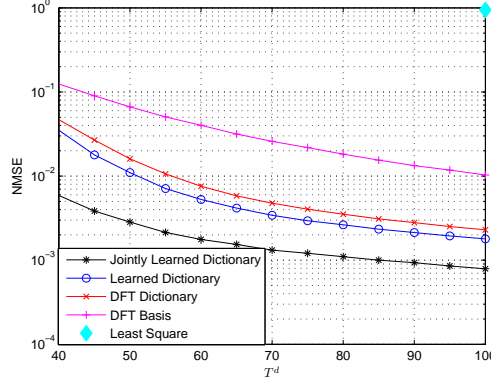


Fig. 8. Normalized mean square error (NMSE) comparison of different sparsifying basis and dictionaries. SNR = 30dB.

complex Gaussian variables with mean 0 and variance 1. It is assumed that the uplink training period $T^u = 10$, and $\rho^u = \rho^d$ such that each antenna at base station transmits the same power as the user terminal. We use Algorithm 2 to perform compressed channel estimation using JDLCM and Algorithm 1 to perform channel estimation using DLCM, as well as using DFT dictionary and DFT basis.

In Fig. 8, it shows that the compressed channel estimation using JDLCM can achieve much better performance than using DLCM and DFT dictionary/basis, and the advantage is most obvious when T^d is small. For example, to achieve $\text{NMSE} = 6 \times 10^{-3}$, JDLCM only needs $T^d = 40$, while DLCM requires $T^d = 57$ and DFT dictionary requires $T^d = 64$. The result demonstrates the advantage of utilizing uplink training to help downlink channel estimation. With moderate number of uplink training ($T^u = 10$ in the simulation), one can greatly reduce downlink training period by using JDLCM.

VI. CONCLUSION

In this paper, we develop a dictionary learning based channel modeling approach, which learns a dictionary from samples of channel measurements. The learned dictionary is able to adapt to the cell characteristics and any array uncertainties, leading to a more accurate and robust channel representation compared to the DFT dictionary and DFT basis. When compressed channel estimation using DLCM is applied, accurate channel estimation is achieved with fewer training symbols saving downlink training overhead compared to those using a predefined

basis/dictionary. Observing the angular reciprocity between the uplink and downlink channel responses, we develop a joint dictionary learning based channel modeling approach which utilizes the reciprocity structure to achieve better downlink channel estimation. Extensive simulation results demonstrate the benefits of representation and estimation of the channel using DLDM and JDLM. The results of the paper show that the important concepts of sparsity and learning from the data can be useful for future communication systems. In particular, as the dimension of the channel vector increases in the future communication systems, sparse representation provides an avenue to deal with the curse of dimensionality. Furthermore, with the availability of more data and computational resources, learning from the data will bring new opportunities to improve performance. As future work, online dictionary learning [16] to deal with slowly changing cell and antenna characteristics as well as quantization effects due to limited feedback resources [48] are under consideration.

REFERENCES

- [1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 186–195, 2014.
- [2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up mimo: Opportunities and challenges with very large arrays," *Signal Processing Mag., IEEE*, vol. 30, no. 1, pp. 40–60, 2013.
- [3] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing the large-scale array regime," *Information Theory, IEEE Transactions on*, vol. 59, no. 10, pp. 6441–6463, 2013.
- [4] X. Rao and V. K. Lau, "Distributed compressive csit estimation and feedback for fdd multi-user massive mimo systems," *Signal Processing, IEEE Transactions on*, vol. 62, no. 12, pp. 3261–3271, 2014.
- [5] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *Communications Magazine, IEEE*, vol. 48, no. 11, pp. 164–174, 2010.
- [6] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.
- [7] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.
- [8] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [9] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave mimo systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, 2014.
- [10] A. M. Sayeed, "Deconstructing multiantenna fading channels," *Signal Processing, IEEE Transactions on*, vol. 50, no. 10, pp. 2563–2579, 2002.
- [11] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

- [12] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5. IEEE, 1999, pp. 2443–2446.
- [13] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003.
- [14] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [15] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in neural information processing systems*, 2006, pp. 801–808.
- [16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.
- [17] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [18] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.
- [19] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 921–928.
- [20] G. G. Raleigh, S. N. Diggavi, V. K. Jones, and A. Paulraj, "A blind adaptive transmit antenna algorithm for wireless communication," in *Communications, 1995. ICC'95 Seattle, 'Gateway to Globalization', 1995 IEEE International Conference on*, vol. 3. IEEE, 1995, pp. 1494–1499.
- [21] P. Zetterberg and B. Ottersten, "The spectrum efficiency of a base station antenna array system for spatially selective transmission," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 3, pp. 651–660, 1995.
- [22] A. J. Paulraj and C. B. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, vol. 14, no. 6, pp. 49–83, 1997.
- [23] Y.-C. Liang and F. P. S. Chin, "Downlink channel covariance matrix (dccm) estimation and its applications in wireless ds-cdma systems," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 222–232, 2001.
- [24] K. Hugl, K. Kalliola, and J. Laurila, "Spatial reciprocity of uplink and downlink radio channels in fdd systems," *Proc. COST 273 Technical Document TD (02)*, vol. 66, p. 7, 2002.
- [25] H. Dai, L. Mailaender, and H. V. Poor, "Cdma downlink transmission with transmit antenna arrays and power control in multipath fading channels," *EURASIP Journal on wireless communications and networking*, vol. 2004, no. 1, pp. 32–45, 2004.
- [26] K. Hugl, J. Laurila, and E. Bonek, "Downlink beamforming for frequency division duplex systems," in *Global Telecommunications Conference, 1999. GLOBECOM'99*, vol. 4. IEEE, 1999, pp. 2097–2101.
- [27] B. K. Chalise, L. Haering, and A. Czylik, "Robust uplink to downlink spatial covariance matrix transformation for downlink beamforming," in *Communications, 2004 IEEE International Conference on*, vol. 5. IEEE, 2004, pp. 3010–3014.
- [28] B. M. Hochwald and T. Mureeta, "Adapting a downlink array from uplink measurements," *Signal Processing, IEEE Transactions on*, vol. 49, no. 3, pp. 642–653, 2001.
- [29] Y. Ding and B. D. Rao, "Compressed downlink channel estimation based on dictionary learning in fdd massive mimo systems," in *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2015, pp. 1–6.

- [30] —, “Channel estimation using joint dictionary learning in fdd massive mimo systems,” in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015, pp. 185–189.
- [31] J. Hoydis, S. Ten Brink, and M. Debbah, “Massive mimo in the ul/dl of cellular networks: How many antennas do we need?” *IEEE Journal on selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, 2013.
- [32] S. M. Kay, *Fundamentals of Statistical Signal Processing: Practical Algorithm Development*. Pearson Education, 2013, vol. 3.
- [33] Z. Gao, L. Dai, Z. Wang, and S. Chen, “Spatially common sparsity based adaptive channel estimation and feedback for fdd massive mimo,” *IEEE Transactions on Signal Processing*, vol. 63, no. 23, pp. 6169–6183, 2015.
- [34] J.-C. Shen, J. Zhang, E. Alsusa, and K. Letaief, “Compressed csi acquisition in fdd massive mimo: How much training is needed?” *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4145–4156, 2016.
- [35] Y. C. Pati, R. Rezaifar, and P. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.
- [36] I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm,” *Signal Processing, IEEE Transactions on*, vol. 45, no. 3, pp. 600–616, 1997.
- [37] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [38] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
- [39] D. P. Wipf and B. D. Rao, “Sparse bayesian learning for basis selection,” *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [40] 3GPP, “Universal Mobile Telecommunications System (UMTS); Spatial channel model for Multiple Input Multiple Output (MIMO) simulations,” 3rd Generation Partnership Project (3GPP), TR 25.996 version 12.0.0 Release 12, Sep 2014.
- [41] A. F. Molisch, A. Kuchar, J. Laurila, K. Hugel, and R. Schmalenberger, “Geometry-based directional model for mobile radio channels principles and implementation,” *European Transactions on Telecommunications*, vol. 14, no. 4, pp. 351–359, 2003.
- [42] A. F. Molisch and F. Tufvesson, “Propagation channel models for next-generation wireless communications systems,” *IEEE Transactions on Communications*, vol. 97, no. 10, pp. 2022–2034, 2014.
- [43] H. L. Van Trees, *Detection, estimation, and modulation theory, optimum array processing*. John Wiley & Sons, 2004.
- [44] H. Rauhut, K. Schnass, and P. Vandergheynst, “Compressed sensing and redundant dictionaries,” *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2210–2219, 2008.
- [45] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall, “Compressed sensing with coherent and redundant dictionaries,” *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.
- [46] P. Tseng, “Convergence of a block coordinate descent method for nondifferentiable minimization,” *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.
- [47] R. Remi and K. Schnass, “Dictionary identification sparse matrix-factorization via-minimization,” *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.
- [48] Z. Zhou, X. Chen, D. Guo, and M. L. Honig, “Sparse channel estimation for massive mimo with 1-bit feedback per dimension,” *arXiv preprint arXiv:1610.03514*, 2016.
- [49] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, “Distributed compressed sensing,” 2005.

- [50] S. Ji, D. Dunson, and L. Carin, “Multitask compressive sensing,” *Signal Processing, IEEE Transactions on*, vol. 57, no. 1, pp. 92–106, 2009.
- [51] Y. Ding and B. D. Rao, “Joint dictionary learning and recovery algorithms in a jointly sparse framework,” in *2015 49th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2015, pp. 1482–1486.
- [52] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 7, pp. 2477–2488, 2005.
- [53] Z. Zhang and B. D. Rao, “Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 5, pp. 912–926, 2011.
- [54] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [55] E. Van Den Berg and M. P. Friedlander, “Probing the pareto frontier for basis pursuit solutions,” *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.